DOCUMENT RESUME

ED 163 089 TM 008 231

AUTHOR Carlson, Alfred B.; And Others
TITLE The Development and Pilot Testing of Criterion Rating Scales. GRE Board Professional Report GREB No. 73-1P.
INSTITUTION Educational Testing Service, Princeton, N.J.
PUB DATE Oct 76
NOTE 77p.; Not available in hard copy due to print quality
AVAILABLE FROM Graduate Record Examinations, Educational Testing Service, Princeton, New Jersey 08541 (free while supplies last)

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS *Behavior Rating Scales; Chemistry; College Entrance Examinations; College Faculty; English; *Evaluation Criteria; Grade Point Average; *Graduate Students; Graduate Study; Higher Education; Predictive Validity; Psychology; Scores; Student Behavior; *Student Evaluation; Test Construction; *Test Reliability; *Test Validity
IDENTIFIERS Graduate Record Examinations; *Rater Reliability

ABSTRACT

Rating scales of the "scaled behavioral expectation" type were developed to measure the constructs of independence and initiative, conscientiousness, enthusiasm, critical facility, teaching skills, research and experimentation, communication, and persistence. The scales were used by faculty in three psychology departments, two chemistry departments, and one English department, who indicated the skills and behavior which were expected of graduate students in their department. The following data were also collected: faculty member's estimate of confidence for each rating; Graduate Record Examinations (GRE) Aptitude and Advanced Test scores; amount of graduate work completed; and undergraduate and graduate grade point average. The scales were found to have only minimal reliability and rather high intercorrelations. Further research on the scales is necessary before they can be used with any confidence. The available data did not suggest any striking relationship between GRE scores and faculty ratings; thus, it is doubtful that such ratings would be useful as criteria in GRE validity studies. (Author/BCP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED163089

TM008 231

THE GRADUATE RECORD EXAMINATIONS

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

GRE

THE DEVELOPMENT AND PILOT TESTING OF CRITERION RATING SCALES

Alfred B. Carlson
Richard R. Reilly
Margaret H. Mahoney
Patricia L. Casserly

GRE Board Professional Report GREB No. 73-1P

October 1976

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Educational Testing Service

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY □ BERKELEY, CALIFORNIA □ EVANSTON, ILLINOIS

The Development and Pilot Testing of Criterion Rating Scales

Alfred B. Carlson
Richard R. Reilly
Margaret H. Mahoney
Patricia L. Casserly

GRE Board Professional Report GREB No. 73-1P

October 1976

Copyright © 1976 by Educational Testing Service. All rights reserved.

Abstract

Rating scales of the "scaled behavioral expectation" type were developed to measure the constructs of independence and initiative, conscientiousness, enthusiasm, critical facility, teaching skills, research and experimentation, communication, and persistence. The scales were used by faculty in three psychology departments, two chemistry departments, and one English department.

The scales were found to have only minimal reliability and rather high intercorrelations. Further research on the scales is necessary before they can be used with any confidence.

The Development and Pilot Testing of Criterion Rating Scales

Obtaining useful measurements of the performance of graduate students has been a persistent concern for researchers in assessing the validity of tests or other instruments in the graduate school context. Lannholm, Marco, and Schrader (1968) discuss the difficulties in assessing the predictive validity of admissions instruments. Reilly (1971) summarizing some of these difficulties stated:

> First, the small sample sizes available at the graduate level make results, especially when several predictors are involved, subject to a considerable degree of error. Second, the fact that students within a given department have gone through an elaborate screening process, and as a result are usually quite homogeneous with respect to most predictor information, often leads to restricted variation in the predictor score distributions. Finally, there is the difficulty of establishing an adequate criterion of graduate school performance. Grade point average (GPA), while it has been the most widely used criterion, has also been the most severely criticized. Perhaps the most important and valid of these criticisms is that the GPA represents only a limited aspect of graduate school performance. (p. 1)

In a later report to the GRE Board Research Committee, Reilly (1974) summarized the results of a two-phase study that was aimed at empirically defining dimensions of graduate student performance. During the first phase of the study, a series of "critical" incidents which reflected unusually effective or unusually ineffective performance was collected from faculty members. A final edited list of 52 incidents was then used as a checklist by the faculty in departments of chemistry, English, and psychology to evaluate the performance of selected students.

The study resulted in the identification, through factor analysis, of eight relatively coherent dimensions of performance represented by clusters of incidents. The eight dimensions were independence and initiative, conscientiousness, enthusiasm, critical facility, teaching skills, research and experimentation, communication, and persistence. A fair degree of consistency existed across the three fields, and the identification of

these factors served as a first step in the development of a usable set of criteria for assessing graduate student performance.

The purpose of the study reported here was to develop a set of "behaviorally anchored" rating scales by which the factors identified in the first phase of the study could be measured and to pilot test the scales. The pilot study was to explore the feasibility of graduate faculty using the developed scales and to determine the psychometric adequacy of the scales. It would also serve as a way of obtaining faculty reaction to specific aspects of the scales.

Rating scales have been the subject of research by psychologists for many years. Although there are many problems with rating scales (Guilford, 1954), their appeal to evaluators and researchers alike is sufficiently great that they continue to be one of the most commonly used techniques of evaluation. The "scaled behavioral expectation" technique developed by Smith and Kendall (1963) is generally recognized as a technique that can effectively overcome many of the problems that seem to be inherent in using rating scales. This technique is designed to provide as much help to the rater as possible in making his judgments. Expected behaviors are used to encourage him to be conscientious; involvement of the raters' peers is intended to maximize communication through use of appropriate terminology and to insure a high degree of content validity.

However, sets of rating scales that have been developed using the scaled behavioral expectation technique tend to require that the rater have a very thorough acquaintance with and knowledge of the ratee. In the context of graduate business school, it was found that many raters (professors) did not know the ratees (students) well enough to make informed ratings on many of the dimensions (Hilton, Kendall, & Sprecher, 1970). For this reason, one of the main purposes of the pilot study was to explore the feasibility of using the scales. Not only was there a question of whether the faculty would know the students well enough to make adequate ratings, but there was also the question about the willingness of departments and faculty to utilize the scales for purposes of research.

A study of common criteria in graduate education by Carlson, Evans, and Kuykendall (1973) suggested that rating scales would be an acceptable criterion measure in many fields. Many departments use rating procedures to select students for financial aid, to determine which students will be encouraged to continue, etc. In such cases, these ratings serve a particular need for the department, and the procedures are developed for a specific purpose. In addition to being perhaps more difficult to complete, a more general set of scales for evaluating graduate students may not be as useful to the department. Also, many faculty members are genuinely concerned about confidentiality of evaluative information, and others are concerned because of recent laws and rulings. Thus, the question of the willingness of faculty and departments to utilize the scales, particularly for research purposes, is a very real one.

Another purpose of the pilot study was to examine the scales from a psychometric point of view--the interrater reliability of each scale, the correlations among the scales, and the correlations between each of the scales and other information available on the students. The interrater reliability of a scale can be thought of as the average correlation between raters when they are rating the same group of students. Ideally, each scale would have a reliability of at least .50, although many rating scales that have been developed do not achieve this level. If this level can be achieved, however, it is possible to obtain a reasonably reliable rating of an individual by averaging the ratings assigned by several raters.

At the same time, the correlations among the scales would ideally be less than .50. The construct that each scale is developed to measure was identified from an analysis of independent factors; if the correlation among the scales is high, then the scale is not adequately measuring the factor construct. Also, if the correlation between two scales is relatively high, then only one of the two scales is necessary since they are measuring essentially the same construct.

Finally, if at least some of the scales are sufficiently reliable and have only small or moderate correlations with each other, the correlation of

these scales with other information on the student should be consistent with logical expectations concerning the meaning of the scale in question. For example, one would expect that the correlation between GRE scores and ratings of "critical facility" would be higher than the correlation between these scores and "persistence" ratings. At the same time, undergraduate grades may well have a moderate or even high correlation with "persistence."

The examination of correlations was a secondary purpose of the pilot study. Participating departments were asked to supply readily available "predictor" data (such as GRE scores or undergraduate grade-point average) and "criterion" data (such as routinely obtained departmental ratings, graduate grade-point average, or prelim scores). These data could provide useful information for increasing the understanding of the "meaning" of the scales.

Method

Scale Development

The approach taken in developing the scales was consistent with methods outlined by Smith and Kendall (1963) in the description of the scaled behavioral expectation technique. Briefly, the steps were as follows:

1. A general definition of each scale was written.

2. A pool of specific behavioral examples for later assignment to the scales was prepared and edited to appear in a common format. The original list of 52 incidents was part of this pool, as were other incidents suggested by faculty respondents during the course of the previous study. Behavioral examples were also culled from those collected for the ATGSB Criterion Study (Hilton, Kendall, & Sprecher, 1970). In addition, a number of new examples were written, many for the mid-ranges of the scales, since the original list of behavioral examples represented only extremes of performance.

3. A pool of 12 judges (four from each discipline) assigned behavioral examples to scales based on their judged relevance to the general definitions in Step 1. Particular attention was paid to disagreement between fields on any of the examples. Only examples for which there was strong agreement among judges were retained.

4. Scale values were assigned to the subset of behavioral examples assigned to each scale. The same pool of judges rated each example in terms of the degree to which they thought that it reflected effective (or ineffective) performance on the continuum defined in Step 1. The distribution of scale values and other information for each behavioral example are given in Appendix A.

5. A final set of behavioral examples selected to anchor each scale was based on two criteria. First, the degree of agreement among judges was considered. Examples where there was lack of agreement in scaled values were eliminated. Second, a set of examples which covered the entire scale continuum was chosen so that only one of two or more examples with the same, or nearly the same, scale values were retained. For these scales, from five to seven examples were considered sufficient to represent anchors over the range of scale values. The final set of rating scales is included in Appendix B, pp. 47-56.

Data Collection

Plans for the data collection included obtaining cooperation from at least three departments within each of the disciplines studied; each department should be large enough to anticipate providing ratings for at least 75 students. An effort was made to restrict student ratees to those who had completed at least two years of graduate study and to have at least two-thirds of the students rated by two faculty members. Within each department, a coordinator was designated and given a set of detailed instructions (see Appendix B).

In addition to ratings by faculty, the following data were collected where possible:

1. Estimates of the confidence a faculty member had for each rating made;
2. GRE Aptitude Test scores;
3. GRE Advanced Test scores;
4. Number of semesters (or quarters) of graduate study;
5. Undergraduate GPA;
6. Graduate GPA, and
7. Any other measures of graduate student performance routinely available (e.g., departmental ratings, class rank, Ph.D. prelim scores).

Analyses

Means, standard deviations, and score ranges were computed for each rating scale to provide information on the extent to which faculty used the full range of the scales. Frequencies of confidence estimates were also tabulated.

Scale reliabilities were estimated for the subsample of students with two ratings for each scale through an analysis of variance procedure described by Winer (1962, p. 126). Reliability of the average of two (or more) raters is estimated as

$$1 - \frac{\text{MEAN SQUARE WITHIN RATEES}}{\text{MEAN SQUARE BETWEEN RATEES}} .$$

The reliability for one rater was estimated by the Spearman-Brown formula (Guilford, 1954, p. 354).

The correlations among the scales and correlations between scales and other data were computed for each department. (The "second" rating for each student who had two ratings was not included in this analysis.) Since most of these sample sizes were quite small, a significance test was computed for each of these correlations. The significance level chosen was .01 because of the large number of coefficients being considered. Data were pooled within discipline to examine the correlations among the scales. Because the other

data collected were so scattered, it was decided to compute correlations with these variables on a departmental basis and to present these correlations only for those variables for which sample sizes were minimally adequate.

Departmental and Faculty Cooperation

Initial contacts with departments were generally made through the graduate dean's office. The purpose of the project was explained, and the most expeditious manner of soliciting the cooperation of the appropriate departments was discussed. With the dean's approval, department chairmen were then approached. As mentioned previously, original plans called for participation of approximately nine departments, three each in the fields of chemistry, psychology, and English. Seven institutions were contacted. The final distribution of participants in the study is shown in Table 1.

Although cooperation was obtained from department chairmen, there was, of course, no guarantee that faculty would agree to participate. Procedures for gathering data were purposely informal with the hope that this would maximize the amount of information collected from the various departments. However, considerable faculty resistance was encountered, and the samples obtained were smaller than anticipated. Suggested reasons for the high rate of refusals are multiple, complex, and varied among departments; therefore, no attempt has been made to order them by frequency of importance. The reasons appear to include: (a) faculty time pressures, (b) a general dislike of rating scales, (c) lack of familiarity with a student's work, (d) the feeling that these rating scales did not pertain to the kind of evaluation of students that took place in the department, and (e) the general tenor of the times (concern about the Buckley Amendment; resistance to any suggestions of invasion of privacy or contribution to data banks in the era of growing information deposits).

Table 2 summarizes the amount of data actually provided by graduate departments. Psychology was the only area in which reasonable sample sizes

Table 1

Distribution of Participants in the Study, by Department

Department	Number of Departments Contacted	Number Agreeing to Participate	Number Providing Data
Psychology	4	4	3
Chemistry	4	2	2
English	6	2	1

Table 2

Distribution by Department of Ratings and Additional Information Obtained

Department	Total Students Rated	Total Students Rated Twice	GRE Verbal	GRE Quantitative	GRE Advanced	Miller Analogies	Other Routinely Available Measures of Graduate Performance	Number of Quarters/ Semesters of Graduate Study Completed	Graduate GPA	Undergraduate GPA
Psychology							(75)			
A	77	55	13	13	10	46	(6)	75	76	45
B	43	25	33	33	20	--	--	39	43	24
C	36	25	34	34	30	--	--	34	36	36
Total	156	105	80	80	60	46	81	148	155	105
Chemistry										
A	67	26	--	--	--	--	--	--	--	--
B	44	14	14	14	13	--	35	43	43	44
Total	111	40	14	14	13	--	35	43	43	44
English	23	13	12	12	11	--	22	23	23	22

were obtained for the purposes of a reliability analysis. Data on other variables were generally sketchy, particularly for GRE scores. The most consistently provided additional data were graduate and undergraduate GPA. Overall, the totals for psychology and chemistry, though smaller than anticipated, could be considered at least minimally acceptable for the pilot study. For English, however, the data provided are clearly of very limited value, especially for purposes of estimating rating scale reliability.

Results of Data Analyses

A first concern with respect to the usefulness of any set of rating scales is the extent to which the entire range of the scale is utilized by raters. Normally one would expect raters to be on the lenient side with ratings but to make at least several ratings at the lower end of the scale. Tables 3, 4, and 5 present means, standard deviations, and score ranges for each discipline studied. It is clear from these tables that raters do tend to use primarily the higher end of the scales. However, it is also clear from the high and low scores given within each department that most of the range is being utilized. The average rating given in most departments was near 4.0, and the standard deviations for each scale averaged approximately 7/10 of a scale unit.

Because the rating scales represented a common set of variables, it was decided to pool rating scale data across departments within discipline so that more stable estimates of scale intercorrelations and reliability could be derived. Table 6 presents scale intercorrelations and reliabilities for one and two raters for psychology.[1] With the exception of Critical Facility, the scales as used in psychology departments appear to possess at least modest reliability for two raters. Correlations between scales, however, are high considering the level of reliability, suggesting a marked halo effect. Table 7 presents

---

[1]Intercorrelations for Tables 6 through 11 were based on the maximum amount of data available for each computation. Thus, the numbers of cases actually used vary slightly from those reported in Table 2.

Table 3

Means, Standard Deviations and Score Ranges
for Rating Scales in Three Psychology Departments

Scale	Psychology A Mean	S.D.	Score Range High	Low
Communication	3.99	0.74	5.00	1.50
Conscientiousness	4.12	0.85	5.00	2.00
Critical Facilty	4.16	0.66	5.00	1.50
Independence and Initiative	4.10	0.82	5.00	1.75
Involvement	4.11	0.76	5.00	1.50
Persistence	4.08	0.70	5.00	1.75
Research	4.02	0.77	5.00	1.25
Teaching	4.07	0.67	5.00	1.00

Scale	Psychology B Mean	S.D.	Score Range High	Low
Communication	3.65	0.70	4.75	1.50
Conscientiousness	3.89	0.81	5.00	2.00
Critical Facility	3.80	0.75	5.00	1.50
Independence and Initiative	3.88	0.83	5.00	1.50
Involvement	3.84	0.75	5.00	2.25
Persistence	3.93	0.70	5.00	1.25
Research	3.62	0.89	4.75	1.25
Teaching	4.05	0.56	5.00	3.00

Scale	Psychology C Mean	S.D.	Score Range High	Low
Communication	3.92	0.57	5.00	1.75
Conscientiousness	4.11	0.87	5.00	1.50
Critical Facility	3.91	0.92	5.00	1.00
Independence and Initiative	3.87	0.93	5.00	1.25
Involvement	3.85	0.88	5.00	1.25
Persistence	4.03	0.69	5.00	2.25
Research	3.83	0.81	5.00	1.00
Teaching	3.96	0.84	5.00	1.50

Table 4

Means, Standard Deviations, and Score Ranges for Rating Scales in Two Chemistry Departments

Scale	Chemistry A Mean	S.D.	Score Range High	Low
Communication	3.62	0.68	4.50	1.75
Conscientiousness	3.91	0.80	5.00	2.00
Critical Facility	3.93	0.53	5.00	2.25
Independence and Initiative	3.63	0.97	5.00	1.25
Involvement	3.66	0.78	5.00	2.25
Persistence	3.99	0.72	5.00	1.75
Research	3.81	0.64	5.00	1.75
Teaching	3.88	0.73	5.00	1.56

Scale	Chemistry B Mean	S.D.	Score Range High	Low
Communication	3.86	0.77	4.75	1.00
Conscientiousness	4.07	0.96	5.00	1.00
Critical Facility	4.12	0.62	5.00	1.50
Independence and Initiative	4.18	0.69	5.00	1.75
Involvement	4.10	0.71	5.00	1.25
Persistence	4.09	0.74	5.00	1.75
Research	4.02	0.48	5.00	3.00
Teaching	4.08	0.84	5.00	1.75

Table 5

Means, Standard Deviations, and Score Ranges for Rating Scales in One English Department

Scale	English			
	Mean	S.D.	Score Range High	Low
Communication	4.07	0.87	4.75	1.55
Conscientiousness	4.38	0.79	5.00	2.00
Critical Facility	4.31	0.81	5.00	1.47
Independence and Initiative	4.25	0.63	5.00	2.13
Involvement	4.14	0.75	4.75	1.63
Persistence	4.25	0.75	5.00	1.25
Research	4.22	0.79	5.00	1.59
Teaching	4.19	0.91	5.00	1.74

Table 6

Rating Scale Intercorrelations and Reliabilities

Pooled Over Three Psychology Departments

Scale	Communi-cation	Conscien-tiousness	Critical Facility	Indep. and Initiative	Involve-ment	Persis-tence	Research	Teaching
Communication		.45	.49	.44	.49	.32	.47	.53
Conscientiousness			.44	.55	.54	.46	.62	.35
Critical Facility				.48	.43	.35	.48	.34
Independence and Initiative					.64	.62	.78	.34
Involvement						.56	.58	.44
Persistence							.61	.31
Research								.37
Teaching								
Reliabilities								
For one rater	.25	.47	.10	.25	.33	.36	.30	.25
For two raters	.40	.64	.19	.40	.50	.53	.46	.40

Table 7

Rating Scale Intercorrelations and Reliabilities

Pooled Over Two Chemistry Departments

Scale	Communi-cation	Conscien-tiousness	Critical Facility	Indep. and Initiative	Involve-ment	Persis-tence	Research	Teaching
Communication		.54	.55	.40	.51	.39	.61	.54
Conscientiousness			.45	.44	.47	.44	.65	.53
Critical Facility				.42	.59	.39	.55	.36
Independence and Initiative					.48	.49	.71	.28
Involvement						.56	.58	.38
Persistence							.62	.29
Research								.42
Teaching								
Reliabilities								
For one rater	.50	.38	.37	.24	.17	.00	.27	.43
For two raters	.67	.55	.54	.39	.29	.00	.42	.60

reliabilities and intercorrelations for chemistry departments. Six of the scales have at least modest reliability for two raters. The persistence scale was completely unreliable as estimated from these data, and the reliability for the involvement scale was very low (.17). Again, the intercorrelations among scales are high considering the level of reliability.

Table 8 (based on a very small sample, as noted previously) presents correlations between scales and reliabilities for the one English department that provided data. Allowing for rater sampling variance, the pattern is consistent with that observed in the other two disciplines. That is, with one or two exceptions scales have at least moderate reliabilities for two raters and scale intercorrelations are high.

Table 9 presents correlations between rating scales and selected variables for each of the three psychology departments. The largest number of correlations significantly greater than zero are in the table for Psychology Department A. All of the correlations between average prelim scores and rating scales were significant. Aside from that set of correlations, however, few of the remaining relationships were significant.

Only one chemistry department was able to provide additional data. The results, presented in Table 10, are unimpressive. The only significant relationship out of 48 computed is a negative correlation between rank-in-class and ratings on the teaching scale. In the English department (see Table 11), significant correlations were obtained between three rating scales and graduate grade-point average, but none of the other correlations reached significance.

A final set of data is presented in Table 12. The percentages of raters expressing various levels of confidence for each rating suggest that an overwhelming majority of raters felt at least "fairly confident" in their ratings.

Discussion and Conclusions

In the authors' judgment, the development of the rating scales went very well, and all of the usual criteria for successful development of scales of this sort were clearly met.

Table 8

Rating Scale Intercorrelations and Reliabilities
in One English Department

Scale	Communi-cation	Conscien-tiousness	Critical Facility	Indep. and Initiative	Involve-ment	Persis-tence	Research	Teaching
Communication		.55	.66	.68	.53	-.22	.93	.87
Conscientiousness			.62	.64	.53	.50	.41	.39
Critical Facility				.46	.06	.27	.70	.56
Independence and Initiative					.66	.02	.62	.58
Involvement						-.14	.38	.45
Persistence							-.26	-.22
Research								.81
Teaching								
Reliabilities								
For one rater	.34	.00	.16	.61	.24	.34	.26	.42
For two raters	.51	.00	.28	.76	.39	.51	.41	.59

Table 9

Correlations Between Rating Scales and Selected Variables in Three Psychology Departments

Psychology A

Scale	Undergraduate GPA	Graduate GPA	GRE Verbal	GRE Quantitative	GRE Advanced	Average Prelim Score	Miller Analogies
Communication	.01	.10	-.22	-.42	.57	.31*	.09
Conscientiousness	.43	.36*	-.13	-.02	-.28	.32*	.07
Critical Facility	.31	.09	-.19	-.03	.10	.23	.07
Independence & Initiative	.17	.17	.03	.04	-.04	.32*	.16
Involvement	.13	.26*	.00	-.29	.15	.26*	.10
Persistence	.17	.20*	-.06	-.24	-.36	.33*	.01
Research	.25	.28*	-.09	.08	-.02	.36*	.11
Teaching	.10	.08	-.14	-.19	.11	.42*	.08

Psychology B

Scale	Undergraduate GPA	Graduate GPA	GRE Verbal	GRE Quantitative	GRE Advanced
Communication	.24	.12	.26	-.03	.03
Conscientiousness	.44	.18*	.07	-.28	-.35
Critical Facility	.19	.40*	-.04	-.11	.04
Independence & Initiative	-.05	.27	.04	-.03	-.10
Involvement	.14*	.28	.22	.07	.11
Persistence	.48*	-.01	.10	.13	-.03
Research	.16	.12	-.00	-.11	-.00
Teaching	.22	.30	.12	.18	-.18

Psychology C

Scale	Undergraduate GPA	Graduate GPA	GRE Verbal	GRE Quantitative	GRE Advanced
Communication	.13	.42*	.22	.05	.31
Conscientiousness	.23	.31	-.03	-.01	.11
Critical Facility	.19	.21	-.06	.04	-.00*
Independence & Initiative	.23	.30	.31	.26	.45*
Involvement	.04	.10	.02	.24	.21
Persistence	.08	.22	.24	.16	.20
Research	.35	.34	.25	.03	.33
Teaching	-.06	-.21	-.17	.28	.02

*Significant at .01 level.

Table 10

Correlations Between Rating Scales and Selected Variables in One Chemistry Department

Scale	Undergraduate GPA	Graduate GPA	GRE Verbal	GRE Quantitative	GRE Advanced	Rank-in-Class
Communication	-.04	.23	.43	-.24	.23	-.02
Conscientiousness	.16	.26	.51	-.24	-.14	-.08
Critical Facility	.14	.33	.32	.04	.33	.01
Independence & Initiative	-.17	-.01	.24	.06	.17	-.14
Involvement	.19	.21	-.12	.01	-.42	-.05
Persistence	.10	-.03	-.35	-.02	-.54	-.17
Research	-.28	.08	.01	-.03	-.20	-.03
Teaching	-.01	-.01	.27	-.49	-.10	-.44*

*Significant at .01 level.

Table 11

Correlations Between Rating Scales and Selected Variables in One English Department

Scale	Undergraduate GPA	Graduate GPA	GRE Verbal	GRE Quantitative	GRE Advanced	Master's GPA
Communication	.18	.55*	.10	.28	.41	.25
Conscientiousness	-.04	.19	-.11	.11	.25	.08
Critical Facility	.02	.16	.16	.32	.16	.27
Independence & Initiative	.10	.55*	.22	.28	.32	.09
Involvement	-.04	.36	-.34	-.06	.32	-.02
Persistence	-.40	-.38	-.26	-.28	-.39	-.27
Research	.27	.58*	.18	.39	.49	.29
Teaching	.16	.42	.17	.16	.45	.33

*Significant at .01 level.

Table 12

Percentages of Faculty Indicating Various Levels of Confidence in Ratings

Scale	Psychology			Chemistry			English		
	Very Confident	Fairly Confident	Not Very Confident	Very Confident	Fairly Confident	Not Very Confident	Very Confident	Fairly Confident	Not Very Confident
Communication	63	27	0	51	42	7	77	23	0
Conscientiousness	62	37	1	54	40	6	83	17	0
Critical Facility	64	36	0	50	43	7	78	22	0
Independence and Initiative	59	41	0	54	38	8	91	9	0
Involvement	55	44	1	49	42	9	74	22	4
Persistence	58	41	1	53	41	6	74	17	9
Research	62	36	1	52	40	8	78	17	4
Teaching	56	38	6	45	40	15	55	41	4

23

The difficulty in obtaining cooperation from the planned number of departments was greater than anticipated. (In retrospect, there is a suggestion that had the study been directed toward validation of the GRE, cooperation would have been somewhat better.) Faculty members in all of the departments contacted indicated some degree of being pressed for time. Clearly, many felt that the research was of little direct value to them or to graduate education in their field. Department chairmen and graduate deans seemed more concerned with the problem addressed in this study than did the faculty. The difficulties encountered suggest that widespread acceptance of scales of this type as criterion instruments will not come easily.

Thus, the amount, and perhaps, to some extent, the quality of the data collected for the pilot study were less than originally anticipated. However, the faculty who did cooperate used effectively the entire range of the scales and felt at least fairly confident in the ratings that they made. Those individuals who provided comments and suggestions, in addition to ratings, generally felt that one or more of the scales were not measuring a single construct. In the expected behaviors provided they saw evidence of two constructs. This is clearly a problem and suggests that some of the scales need additional work.

The obtained reliabilities of the scales for one rater are, in most cases, not acceptable. This, in and of itself, does not clearly mitigate against use of the scales. The reliability for two raters is reasonable for many of the scales; in situations where three to five ratings (perhaps student and faculty) could be obtained for each student, the reliability of the average rating would probably be more than adequate.

The correlations among the scales are quite high. This suggests that a large halo effect is operating or that the scales are measuring less than eight separate constructs. The patterns of correlations with other variables provide little help in this regard. If one takes the statistical significance testing seriously, few correlations can be considered as other than zero. The correlation of the scales with average prelim scores in Psychology

Department A (probably the best alternative criterion data obtained) supports the scales as being valid but the uniformity of the level of correlation also supports the interpretation of the correlations among scales as a halo problem.

Clearly, more work needs to be done on the scales before they can be used in an operational context. Under the right set of circumstances, however, it is felt that scales such as these could be used. One of the most direct methods of increasing the reliability of ratings is to increase the number of raters. In light of the results of this study it does not seem feasible to collect three or four faculty ratings, but it might be possible to collect, say, two faculty and two peer ratings. When one considers that the usual yearly grade-point average is based on a total of at least 8 or 10 separate "ratings" of performance it does not seem unreasonable to require at least 3 or 4 ratings per student.

Another approach which might prove more profitable is to involve faculty within each department in at least one stage of scale development. One strategy would be to present a department with the general definitions for the scales used in the present study and allow faculty to assign the specific behavioral anchors. Faculty could generate their own behavioral examples or could select examples from a pool which would be provided. This approach would have the advantage of adaptability to the particular characteristics of the department and might generate a greater degree of interest and commitment on the part of faculty.

One question which is undoubtedly of much interest to those involved with the GRE Testing Program is whether GRE scores can be used to predict ratings. Unfortunately, the data collected do not really provide an answer to this question. The data which were available, however, certainly do not suggest any striking relationship between GRE scores and ratings. This is not unreasonable in view of the fact that the rating scales were designed to be multidimensional and appear on the face to measure factors other than usual academic aptitude variables which are tapped by GRE scores. This

suggests that such rating scales would be more useful in a research context where a variety of predictors (e.g., biographical data, interest scores, etc.) were included than as criteria in GRE validity studies.

References

Carlson, A. B., Evans, F. R., & Kuykendall, N. M. The feasibility of common criterion validity studies of the GRE (ETS RM 73-16). Princeton, N. J.: Educational Testing Service, 1973. (Also GRE Board Professional Report GREB 71-1P and ERIC Accession No. ED 097 367.)

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Hilton, T. L., Kendall, L. M., & Sprecher, T. B. Performance criteria in graduate business study. Parts I and II: Development of rating scales, background data and pilot study (ETS RB 70-3). Princeton, N. J.: Educational Testing Service, 1970.

Lannholm, G. V., Marco, G. L., & Schrader, W. B. Cooperative studies of predicting graduate school success (GRE Special Report Number 68-3). Princeton, N. J.: Educational Testing Service, August 1968.

Reilly, R. R. Critical incidents of graduate student performance (GREB Technical Memorandum No. 1). Princeton, N. J.: Educational Testing Service, 1971. (Also GRE Board Research Report GREB 70-5R.)

Reilly, R. R. Factors in graduate student performance (ETS RB 74-2). Princeton, N. J.: Educational Testing Service, 1974. (Also GRE Board Professional Report GREB 71-2P.)

Smith, P. C., & Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

APPENDIX A

SUMMARY STATISTICS FOR BEHAVIORAL EXAMPLES

The summary statistics for all the behavioral examples included in the study are presented in the following tables (A-1 through A-16). The behavioral examples are grouped by the eight scales. The odd-numbered tables present statistics for behavioral examples used as anchors in the final rating scales. The even-numbered tables present statistics for the remaining behavioral examples.

Each table shows the ratio of judges who assigned each behavioral example to that scale in the initial sort (Step 3), and the mean value of the example as assigned in Step 4. The distribution of scale values is also given in each table.

The frequency distributions of values assigned to the behavioral examples in a few cases do not add up to 12. This reflects the fact that, for reasons unknown, a judge chose not to assign a value to the example even though it was part of the scale.

In the column showing the ratios of judges assigning behavioral examples to the final scale, one will notice that the denominators are not always 12. A ratio of 10/11 for a particular behavioral example means that only 11 of the 12 judges chose to assign that item to any scale at all, but of those judges who did assign the item to a scale, 10 of them agreed that it belonged on the scale to which it was assigned.

Table A-1

Distribution of Scale Values for Behavioral Examples Used as Anchors

Scale: COMMUNICATION

	ITEM			Frequency Distribution									Ratio[2]
	I would expect this student to:	CI[1]	$\bar{X}$	1	1.5	2	2.5	3	3.5	4	4.5	5	
222	present a report at a regional convention which would be well received for its humor and style.		4.5						1	5		6	11/12
206	submit written work that would be interesting and no trouble to read.		3.9					2	1	8		1	11/11
209	usually be understood when he speaks, but fail to make points in written communication understandable.		2.8		1	3	1	7					12/12
215	prepare oral and written communications seemingly unaware that the backgrounds of his listeners or readers may not be the same as his own.		2.2			9	1	2					9/10
202	take ten pages to complete an assigned paper which should be completed in two pages.		1.9	2		9	1						10/12
218	present ideas in a seminar, paper, or test in a poorly organized and disjointed fashion.	X	1.4	7		5							12/12

[1] Item from "critical incidents" study

[2] Ratio of judges assigning behavioral examples to final scale

Table A-2

Distribution of Scale Values for Additional Behavioral Examples
Assigned to Scales but NOT Used as Anchors

Scale: COMMUNICATION

	ITEM		Frequency Distribution									Ratio[1]
	I would expect this student to:	$\tilde{X}$	1	1.5	2	2.5	3	3.5	4	4.5	5	
210	display an unusually accurate and sensitive choice of words in speaking and writing.	5.0							3	1	8	10/12
214	handle a difficult topic with considerable skill when presenting a paper.	4.7							4	1	7	9/12
205	present ideas in a forceful way, giving an impression of well-thought-out and independent views.	4.6					1	1	4		6	10/11
221	articulately defend his position and ideas.	4.4					1	1	7		3	10/11
213	adopt and maintain a witty and urbane tone when appropriate.	4.1					2	1	5	1	3	11/12
212	deal effectively with the denotive and connotative values of words.	4.1					5	1	1		5	11/12
224	present written and oral reports which would be organized and have clear introductions and logical conclusions.	4.0					3		6		3	11/11
211	deal effectively with the lexicon and grammar of the language.	4.0					2	1	6		3	12/12
203	prepare reports which are clear and well-organized, oral or written	4.0									12	11/12
223	use jargon to the point where his papers can only be read by a specialist in his particular area.	2.2		1	5	1	4					12/12
220	submit papers which are extremely verbose.	1.9	2		9		1					9/11
216	use a phrase such as "You Know", in no less than every other sentence when speaking in class.	1.6	5	1	5		1					11/11
204	present material in a disorganized fashion so that an undue amount of time is needed to understand the points being made.	1.6	5		7							11/12
201	be unable to state in an unambiguous way the key definitions and main points covered by the subject matter.	1.4	7	1	4							9/11
219	display an inability to write competently.	1.3	8		4							9/11
208	talk freely but say nothing constructive; to use big words, often erroneously; to be almost impossible to follow.	1.1	11		1							11/12
217	submit a report which is replete with grammatical errors.	1.0	11	1								10/12
207	submit work which is full of mistakes, incomplete sentences and misspellings; almost impossible to follow the line of thought.	1.0	12									10/10

[1] Ratio of judges assigning behavioral examples to final scale

Table A-3

Distribution of Scale Values for Behavioral Examples Used as Anchors

Scale: CONSCIENTIOUSNESS

	ITEM	CI[1]	$\bar{X}$	Frequency Distribution 1	1.5	2	2.5	3	3.5	4	4.5	5	Ratio[2]
	I would expect this student to:												
802	return with a carefully annotated bibliography after having been assigned to track down a number of references.		4.8							3		9	5/12
806	have the reputation amongst both the faculty and his fellow students for doing what he says he will do.		4.4							6	2	4	11/12
801	seldom miss deadlines.		3.7					4		6		1	11/11
808	do each assigned literature search but each would be incomplete.		2.3			8	1	3					9/12
805	fail on one or more occasions to complete a major assignment on time.	X	1.9	2		7	1	2					9/12
812	exhibit carelessness with laboratory equipment.	X	1.6	5		7							10/12
803	miss class repeatedly without contacting the instructor.		1.2	10		2							9/10

[1] Item from "critical incidents" study

[2] Ratio of judges assigning behavioral examples to final scale

Table A-4

Distribution of Scale Values for Additional Behavioral Examples
Assigned to Scales but NOT Used as Anchors

Scale: CONSCIENTIOUSNESS

	ITEM	$\bar{X}$	Frequency Distribution 1	1.5	2	2.5	3	3.5	4	4.5	5	Ratio[1]
	I would expect this student to:											
807	handle even the most menial assignment (e.g., paper grading) with care and responsibility.	4.5					1		4		7	10/12
809	have a reputation for failing to appear for one or two professional appointments each year though he comes on time to most such appointments.	2.4	1		5	1	5					10/10
810	submit a report which is incomplete.	1.9	2		9		1					9/11
811	fail to do background reading for a research project.	1.5	6		6							10/12
814	begin to look for the appropriate references a day before he is to report at a seminar.	1.5	7		4		1					11/12
813	let assignments slide then, either to submit a hastily prepared report or to submit the report well past the deadline.	1.4	7		5							11/11
804	fail to appear for a symposium in which he was to be a participant without warning the other participants.	1.0	12									9/11

[1]Ratio of judges assigning behavioral examples to final scale

Table A-5

Distribution of Scale Values for Behavioral Examples Used as Anchors

Scale: CRITICAL FACILITY

	ITEM: I would expect this student to:	CI[1]	$\bar{X}$	Frequency Distribution 1	1.5	2	2.5	3	3.5	4	4.5	5	Ratio[2]
714	make a perceptive analysis and evaluation of a difficult text in his field of specialization.		4.8					1		1		10	10/12
706	offer well founded qualifications to instructor's statements in class in a discerning, constructive way.		4.5							5	1	6	10/12
711	ask questions which are always relevant and usually perceptive.	X	4.0					1		10		1	10/11
710	be sensitive to faculty evaluations and suggestions; he could incorporate suggestions once his deficiencies were pointed out.		3.8					4		7		1	10/12
709	successfully identify a problem with another student's research but to be extremely harsh in his criticism.		2.9			4		7		1			9/10
713	often be unable to consider new ideas objectively because of strongly held prejudices.	X	1.5	5	2	5							11/12
701	be unreceptive to new ideas and proposals even in situations where previous methods had proven inadequate.		1.3	9		3							11/12

[1] Item from "critical incidents" study

[2] Ratio of judges assigning behavioral examples to final scale

Table A-6

Distribution of Scale Values for Additional Behavioral Examples
Assigned to Scales but NOT Used as Anchors

Scale: CRITICAL FACILITY

	ITEM		Frequency Distribution									Ratio[1]
	I would expect this student to:	$\bar{X}$	1	1.5	2	2.5	3	3.5	4	4.5	5	
715	consistently offer well founded and constructive criticism of other students' presentations.	4.7							3	1	8	11/12
718	examine carefully all authors' premises and frames of reference before accepting conclusions.	4.6							4	1	7	9/12
716	display an openness to evaluation and criticism of his work by others.	4.5					1		4		7	10/11
705	detect inconsistencies in the position taken by a professor in a critical, significant classroom lecture.	4.2					2		5	2	3	9/12
703	back a position objectively and without defensiveness in a discussion on the merit of an issue.	4.1					2		6	1	3	9/11
720	adopt a critical position proper and appropriate to the work under discussion.	4.0					3	1	3	2	3	10/12
719	be able to understand the tone of a work and its underlying assumptions.	4.0					3		5	2	2	9/12
708	generate and adequately support a critical generalization.	4.0					4		4		4	9/12
707	display ability to shift his point of view during a debate or discussion	3.9					5		2		4	9/12
704	assess the limitations of theories and principles from a discipline when they are applied to a specific situation.	3.8			1		5		2		4	10/12
717	be very facile and articulate, but his interpretations would be almost always inappropriate.	1.6	5		7							8/11
712	make exaggerated claims for the relevance or importance of his particular speciality and be unable or unwilling to deal with alternative approaches.	1.4	8		3	1						9/10
702	make unwarranted assumptions based on erroneous information.	1.3	10		2							10/12

[1]Ratio of judges assigning behavioral examples to final scale

Table A-7

Distribution of Scale Values for Behavioral Examples Used as Anchors

Scale: INDEPENDENCE & INITIATIVE

	ITEM			Frequency Distribution									Ratio[2]
	I would expect this student to:	CI[1]	$\bar{X}$	1	1.5	2	2.5	3	3.5	4	4.5	5	
312	learn an important research skill on his own.	X	4.8							2	2	8	9/12
315	display an ability to formulate problems or issues suggested by the material under study rather than the mediation of his professors.		4.5							5	1	6	9/11
303	ask questions and seek information beyond the material in the text or lecture.		3.8					3		8		1	9/12
302	develop a list of several appropriate topics for an assigned research paper, but not to choose one from the list until urged to do so by the professor after the deadline has passed.		2.2	1		7	2	2					9/10
301	depend upon his collaborator for the suggestion of a topic, definition of the problem, and direction of the work in a joint project.		2.0	2		8	1	1					10/12
314	not respond to suggestions or supervision unless the instructor pursues; have to be prodded and pushed and pulled along.		1.2	10		2							8/12

[1] Item from "critical incidents" study

[2] Ratio of judges assigning behavioral examples to final scale

Table A-8

Distribution of Scale Values for Additional Behavioral Examples
Assigned to Scales but NOT Used as Anchors

Scale: INDEPENDENCE & INITIATIVE

	ITEM		Frequency Distribution									Ratio[1]
	I would expect this student to:	$\bar{X}$	1	1.5	2	2.5	3	3.5	4	4.5	5	
307	take on challenging or "not yet officially approved" problems or issues.	4.7							4	1	7	12/12
304	do independent reading or research to check the validity of interpretations or evaluations which differ from those of his instructor or classmates.	4.7					1		1	1	9	9/12
313	become more proficient in a useful outside field under his own initiative.	4.6							4	2	6	11/12
309	obtain a copy of an unpublished research report or paper relevant to his own work through correspondence.	4.3					4		1		7	5/12
306	seek further information independent of class assignments, when a professor raised a question in class, in an attempt to answer the question more extensively.	4.3						1	7	1	3	9/12
308	regularly spend time looking through the appropriate journals when these are not parts of assignments.	4.2					2		5	1	4	8/12
316	familiarize himself with the resources of neighboring libraries and special collections.	4.0					3		6		3	8/12
305	need an instructor's help in finding a topic when required to prepare an assignment on a self-selected topic	2.0	2		7	1	2					10/11
311	constantly seek help from faculty on trivial matters.	1.0	11	1								10/11
310	be heavily dependent on direction from faculty and to appear unable to undertake any independent investigations.	1.0	11	1								11/12

[1] Ratio of judges assigning behavioral examples to final scale

Table A-9

Distribution of Scale Values for Behavioral Examples Used as Anchors

Scale: INVOLVEMENT

	ITEM	CI[1]	$\bar{X}$	Frequency Distribution									Ratio[2]
				1	1.5	2	2.5	3	3.5	4	4.5	5	
	I would expect this student to:												
603	be elected as an officer in a regional or local professional society.[3]		4.7						1	2		9	12/12
617	become quickly and enthusiastically involved in a project.	X	4.5							6		6	10/11
605	display concern and interest in work being conducted by faculty.		4.0					2	1	7		2	8/12
615	attend departmental seminars but to neither participate in the discussions nor volunteer to present a report.		2.7			3	1	8					10/12
611	not attend a meeting, either local or national, of the appropriate professional society.		1.8	4	1	5		2					11/11
616	avoid challenging courses or work.	X	1.4	7		5							9/11

[1]Item from "critical incidents" study

[2]Ratio of judges assigning behavioral examples to final scale

[3]This item was placed incorrectly as an anchor (at approximately 3.0) in the rating scale booklet. Its mean indicates that it should have been placed near the top of this scale (at 4.7).

Table A-10

Distribution of Scale Values for Additional Behavioral Examples
Assigned to Scales but NOT Used as Anchors

Scale: INVOLVEMENT

	ITEM		Frequency Distribution									Ratio[1]
	I would expect this student to:	$\bar{X}$	1	1.5	2	2.5	3	3.5	4	4.5	5	
602	give up a vacation to co-author a paper with a faculty member.	5.0									12	11/12
613	give an original paper at a convention or meeting sponsored by a scholarly society.	4.6					1	1	1	1	8	9/12
614	display a genuine interest in and commitment to his field in informal discussions with faculty.	4.5					1		4	1	6	11/12
601	wait for the instructor outside the classroom after a particularly good lecture with an invitation to join in a further discussion of the subject with a few others.	4.4							7	1	4	9/10
609	appear to be very interested in meeting colleagues, as at conventions.	4.0					2	1	6		3	10/11
608	hold membership in the appropriate professional society or association.	3.5				1	6		3		2	9/11
604	be a student member of a national professional society but not an active participant.	3.1					10	1	1			11/11
607	hold no membership in any professional society and to not appear interested in doing so.	1.6	4	1	6	1						11/11
610	attend seminars only when required to do so.	1.4	7	1	4							9/11
606	become distracted by non-academic, non-professional interests.	1.3	7		3	1			1			10/12
612	seldom, if ever, engage in informal contacts with faculty or fellow graduate students.	1.1	10	1	1							10/11

[1] Ratio of judges assigning behavioral examples to final scale

Table A-11

Distribution of Scale Values for Behavioral Examples Used as Anchors

Scale: PERSISTENCE

	ITEM	CI[1]	$\bar{X}$	Frequency Distribution 1	1.5	2	2.5	3	3.5	4	4.5	5	Ratio[2]
	I would expect this student to:												
508	pursue his interest or ideas despite discouraging advice from faculty and to be successful.	X	4.8							2	1	9	10/12
502	try even harder when a problem of supposedly moderate difficulty resists all initial attempts to resolve it.		4.6							5		7	12/12
514	repeatedly ask questions of faculty until he fully understands an issue.	X	4.1					2		7		3	8/12
509	frequently talk about leaving school as soon as a master's program has been completed though he entered graduate school to get a doctorate.		1.9	1	1	9	1						6/10
512	abandon a project after losing a set of preliminary data.		1.3	8	1	3							11/12

[1] Item from "critical incidents" study

[2] Ratio of judges assigning behavioral examples to final scale

Table A-12

Distribution of Scale Values for Additional Behavioral Examples
Assigned to Scales but NOT Used as Anchors

Scale: PERSISTENCE

	ITEM	Frequency Distribution										Ratio[1]
	I would expect this student to:	$\bar{X}$	1	1.5	2	2.5	3	3.5	4	4.5	5	
501	develop a detailed card file on material relevant to a topic even though it required sticking at a demanding, but unexciting task for several days.	4.8							2	1	9	9/12
507	continue to work on research despite one or more setbacks until it was successfully completed.	4.7							4		8	12/12
503	stick to an assignment until completed even if it was of little personal interest.	4.5					1		3	2	6	9/12
505	consult with other research directors about projects that might be available after his first failure in the research laboratory.	2.6	3		2	1	3		2	1		9/12
504	stay away from the laboratory for two weeks or so when several trials for a particular experimental result have failed.	1.9	2	1	7	1	1					11/12
506	investigate community college teaching positions after receiving one C though he has done well in other courses.	1.8	3	1	7	1						7/8
511	complete very few projects.	1.7	6		4		2					6/11
510	conceive ideas for research but to fail to carry out the research.	1.6	5	1	5		1					8/11
513	begin several original research projects at the same time but to fail to complete any of them.	1.4	8	1	2		1					10/12

[1] Ratio of judges assigning behavioral examples to final scale

Table A-13

Distribution of Scale Values for Behavioral Examples Used as Anchors

Scale: RESEARCH

	ITEM	CI[1]	$\bar{X}$	Frequency Distribution 1	1.5	2	2.5	3	3.5	4	4.5	5	Ratio[2]
	I would expect this student to:												
405	develop an original way of handling a research problem.	X	4.8							2	2	8	11/12
410	be familiar with the latest developments in his field.	X	4.3						1	7	1	3	8/12
404	be systematic in his gathering and ordering of data.		3.9					2		9		1	9/11
401	replicate previous research done by others with a carefully-conducted and well-reported experiment.		3.5					5	1	6			10/12
403	conduct a fairly routine and unexciting research project.		2.9			1		11					8/11
407	confine his attention to research matters of minor importance.		2.3	1		6	1	4					9/12
411	be unable to effectively apply a particular research technique.	X	1.6	4	2	6							12/12
409	attempt to carry out poorly planned research.	X	1.3	8	2	1	1						10/12

[1]Item from "critical incidents" study

[2]Ratio of judges assigning behavioral examples to final scale

Table A-14

Distribution of Scale Values for Additional Behavioral Examples
Assigned to Scales but NOT Used as Anchors

Scale: RESEARCH

	ITEM	$\bar{X}$	Frequency Distribution 1	1.5	2	2.5	3	3.5	4	4.5	5	Ratio[1]
	I would expect this student to:											
408	master a difficult research technique in an unusually short period of time.	4.7							3	2	7	10/11
406	deal insightfully with primary sources.	4.3	1						3	1	7	8/11
413	rely too heavily on one research tool in conducting research.	2.3	1	2	4		5					11/12
414	be unable to formulate a testable hypothesis from a theoretical analysis.	1.7	4	1	6		1					9/11
402	conduct research without providing proper controls so that results are questionable.	1.4	6	3	3							8/12
412	be unfamiliar with a major research tool in his field.	1.4	6	3	3							10/11

[1] Ratio of judges assigning behavioral examples to final scale

Table A-15

Distribution of Scale Values for Behavioral Examples Used as Anchors

Scale: TEACHING

	ITEM I would expect this student to:	CI[1]	$\bar{X}$	Frequency Distribution 1	1.5	2	2.5	3	3.5	4	4.5	5	Ratio[2]
103	show imagination and originality in teaching a traditionally dull topic to an undergraduate class.	X	4.9							1		11	10/11
101	help slower students voluntarily.		4.3					1		6		5	9/10
105	take considerable pains to help undergraduates with their work even though his presentation of material in class is frequently poor.		3.0			1	1	8	2				10/12
119	spend class time doing routine exercises.		2.4			6	3	3					10/11
109	complain about having to teach the introductory course.		2.0	2	1	7		2					9/10
115	teach a class in which his students seek help from other instructors.		1.6	4	1	7							11/12
108	badger students and be generally unsympathetic to legitimate request for time extensions or specific help.		1.2	10		2							12/12

[1]Item from "critical incidents" study

[2]Ratio of judges assigning behavioral examples to final scale

C. 62

C. 63

Table A-16

Distribution of Scale Values for Additional Behavioral Examples
Assigned to Scales but NOT Used as Anchors

Scale: TEACHING

	ITEM		Frequency Distribution									Ratio[1]
	I would expect this student to:	$\bar{X}$	1	1.5	2	2.5	3	3.5	4	4.5	5	
102	stimulate great interest and enthusiasm in undergraduate courses in which he is an instructor.	5.0									12	11/12
114	develop provocative and/or imaginative exercises for his class.	4.8							2		10	12/12
118	establish an atmosphere in which students feel free and eager to talk.	4.4					1		5		6	12/12
116	be sensitive to faculty evaluations and suggestions and to change his teaching strategies when deficiencies are pointed out.	4.3					2		4	2	4	10/10
123	identify students with difficulties and make appointments with them.	4.2					2		5	1	4	12/12
104	revise and restructure an entire introductory course.	4.2			1		2		3		6	10/11
117	develop criteria and set up consistent standards by which to evaluate student work.	3.9					5		3		4	12/12
121	schedule office hours for his students and try very hard to keep them.	3.9					5		3	1	3	9/12
112	plan and implement an effect syllabus for a course.	3.8	1				3		3	2	3	11/12
126	ask others for ideas on how to present complex concepts to his students.	3.6			2		3		4	1	2	9/11
120	never miss a class he is to teach but not to announce office hours	2.8			3	2	6		1			9/11
111	know the names of about a third of his students in the class of 25 students which he teaches.	2.3	2		4	2	4					9/11
113	conduct highly structured classes marked by a lack of flexibility and open interchange of ideas.	2.2	1		7	1	3					10/10
125	not cover all the appropriate subject matter in the course he teaches.	2.0	2		8		2					10/11
127	establish rigid criteria and a set of inflexible standards by which to evaluate student work.	1.9	3		7	1	1					10/11
110	come to classes he teaches but have no lesson plan.	1.7	5		5	1	1					10/11
106	be extremely sensitive to student pressure or criticism which would be reflected in the assignment of poor grades.	1.6	7	1	1		3					9/10
122	teach a class which the better undergraduate students consistently cut.	1.5	5	1	6							12/12
124	teach a class in which a large percentage of the students drop the course to switch to another section.	1.5	7		4		1					11/12
107	curry favor with his students; make the demands of the discipline secondary to a desire for popularity.	1.3	9		3							9/11

[1] Ratio of judges assigning behavioral examples to final scale

APPENDIX B

Criterion Rating Scale Study

INSTRUCTIONS FOR DISTRIBUTING AND CODING RATING SCALE BOOKLETS AND RECORDING INFORMATION ON STUDENT RECORD SUMMARY FORMS

The person in charge of gathering information should carry out the following activities:

- Determine which raters will be evaluating which students (an attempt should be made to see that an individual rater is not asked to rate an extremely large number of students).
- See that each professor receives his rating assignments on time.
- Follow-up non-responses at the appropriate time.
- Arrange for rostering test scores and other information requested from students' files.
- See that all information is organized and mailed to ETS as soon as possible.

Rating Scale Booklets

1. To keep information confidential, it is suggested that the data coordinator assign an arbitrary number to each student. A number should also be assigned to each faculty rater. The coordinator should keep a master list of student and faculty identification numbers. When the ratings have been completed, the coordinator should write in the identification numbers from the master list on the covers of the booklets and strike out the students' and raters' names. The coordinator should retain the master list until ETS has received and edited the data. At the completion of the study ETS will destroy the individual data and the master list should be destroyed.

2. It is essential that every student be rated by his or her faculty advisor. In addition, we hope to obtain ratings by a second professor who is familiar with the student's work. It is not necessary that all students be rated by a second professor, but it is essential that we have two ratings for at least two-thirds of the students being rated. The "first rater" should be the student's faculty advisor. His ratings are to be identified by the number "1" to be written in the lower right corner of the cover page of the rating scale booklets. The "second rater" is to be identified by the number "2" in the lower right corner of the cover page of the rating scale booklets.

3. It will not be necessary for the data coordinator to record any of the obtained ratings. ETS will do this.

Student Record Summary Forms

1. Student (No.) — To keep student information confidential, students will be identified only by number. The number assigned to a student for the rating scale booklets should also be used here.

2. Advisor (No.) — The faculty advisor is identified by the number assigned to him for the rating scale booklets.

3. Second Faculty Rater (No.) — The second faculty rater is also identified by the number assigned to him for the rating scale booklets.

4. Sex — self-explanatory.

5. Number of Semesters Enrolled in Graduate School — Record the total number of semesters enrolled, including Spring semester, 1975.

6. Area of Specialization — If student is specializing in more than one area of the field, please list all.

7. GRE Scores — Self-explanatory.

8. Undergraduate GPA — Self-explanatory.

9. Graduate GPA — Furnish the most current overall GPA for the student. If grading system is other than 4.0 = A, please advise us.

10. Core Courses — List course titles and grades.

11. Additional Information — If any other information is available (for example, tests given as evaluative instruments, routine evaluations conducted during graduate work, rank-ordering of students) we would like to have it.

Send the completed rating booklets and student record forms to:

Dr. Richard R. Reilly
Research Psychologist
Developmental Research Division
R-226
Educational Testing Service
Princeton, NJ 08540

Notify ETS when return shipment has been made.

If there are any questions or problems, please call Mrs. Peggy Mahoney at (609) 921-9000, Extension 2383.

On the following pages eight factors which have been identified as important in the performance of graduate students are described by means of a general definition. In addition, various points along the qualitative scale are "anchored" with specific behavioral examples, or incidents of graduate student performance. For each student you rate place a check mark at the point on the scale which in your judgment best describes the student's performance on that particular trait. Also, indicate for each trait the level of confidence you have in your rating.

Student being rated_____

Rated by_____

Communication

General Definition

The ability to transmit ideas and feelings. It includes the degree of organization and precision, it involves the extent of understanding and perception and includes both verbal and nonverbal transmissions.

I WOULD EXPECT THIS STUDENT TO:

5

____ present a report at a regional convention which would be well received for its humor and style.

4

____ submit written work that would be interesting and no trouble to read.

3

____ usually be understood when he speaks, but fail to make points in written communication understandable.

____ prepare oral and written communications seemingly unaware that the backgrounds of his listeners or readers may not be the same as his own.

2

____ take ten pages to complete an assigned paper which should be completed in two pages.

____ present ideas in a seminar, paper, or test in a poorly organized and disjointed fashion.

1

Degree of Confidence

☐ Very Confident ☐ Fairly Confident ☐ Not Very Confident

Conscientiousness

General Definition

The characteristics of carefulness, thoroughness, and commitment to standards. Includes extent of carrying through on commitments even when they are not fully spelled out.

I WOULD EXPECT THIS STUDENT TO:

5 ---- return with a carefully annotated bibliography after having been assigned to track down a number of references.

---- have the reputation amongst both the faculty and his fellow students for doing what he says he will do.

4 ---- seldom miss deadlines.

3 ---- do each assigned literature search but each would be incomplete.

2 ---- fail on one or more occasions to complete a major assignment on time.

---- exhibit carelessness with laboratory equipment.

---- miss class repeatedly without contacting the instructor.

1

Degree of Confidence

☐ Very Confident ☐ Fairly Confident ☐ Not Very Confident

Critical Facility

General Definition

The ability to evaluate the products of others and offer alternative hypotheses, methods or analyses when appropriate. Includes the ability to be persuaded by well reasoned arguments even when they are clearly critical of his own work or position on an issue. The ability to identify problems and structure priorities.

I WOULD EXPECT THIS STUDENT TO:

5

---- make a perceptive analysis and evaluation of a difficult text in his field of specialization.

---- offer well founded qualifications to instructor's statements in class in a discerning, constructive way.

4

---- ask questions which are always relevant and usually perceptive.

---- be sensitive to faculty evaluations and suggestions; he could incorporate suggestions once his deficiencies were pointed out.

3

---- successfully identify a problem with another student's research but to be extremely harsh in his criticism.

2

---- often be unable to consider new ideas objectively because of strongly held prejudices.

---- be unreceptive to new ideas and proposals even in situations where previous methods had proven inadequate.

1

Degree of Confidence

☐ Very Confident ☐ Fairly Confident ☐ Not Very Confident

Independence and Initiative

General Definition

The combination of self confidence and drive. It includes taking responsibility and showing initiative. It may reflect intellectual curiosity and motivation.

I WOULD EXPECT THIS STUDENT TO:

5

---- learn an important research skill on his own.

---- display an ability to formulate problems or issues suggested by the material under study rather than the mediation of his professors.

4

---- ask questions and seek information beyond the material in the text or lecture.

3

develop a list of several appropriate topics for an assigned research paper but not to choose one from the list until urged to do so by the professor
---- after the deadline has passed.

2 ---- depend upon his collaborator for the suggestion of a topic, definition of the problem, and direction of the work in a joint project.

---- not respond to suggestions or supervision unless the instructor pursues; have to be prodded and pushed along.

1

Degree of Confidence

☐ Very Confident ☐ Fairly Confident ☐ Not Very Confident

Involvement

General Definition

The degree of participation and activity in both formal and informal contexts related to the discipline. The extent to which interest and enthusiasm is exhibited and maintained. Could be interpreted as the extent to which a commitment to the field has been made.

I WOULD EXPECT THIS STUDENT TO:

5

---- become quickly and enthusiastically involved in a project.

4 ---- display concern and interest in work being conducted by faculty.

3 ---- be elected as an officer in a regional or local professional society.

---- attend departmental seminars but to neither participate in the discussions nor volunteer to present a report.

2 ---- not attend a meeting, either local or national, of the appropriate professional society.

---- avoid challenging courses of work.

1

Degree of Confidence

☐ Very Confident ☐ Fairly Confident ☐ Not Very Confident

Persistence

General Definition

The characteristic of continuing to pursue a task or idea despite criticism or setbacks.

I WOULD EXPECT THIS STUDENT TO:

5 ---- pursue his interest or ideas despite discouraging advice from faculty and to be successful.

---- try even harder when a problem of supposedly moderate difficulty resists all initial attempts to resolve it.

4 ---- repeatedly ask questions of faculty until he fully understands an issue.

2 ---- frequently talk about leaving school as soon as a master's program has been completed though he entered graduate school to get a doctorate.

---- abandon a project after losing a set of preliminary data.

1

Degree of Confidence

☐ Very Confident ☐ Fairly Confident ☐ Not Very Confident

Research

General Definition

Concerned with curiosity coupled with the desire to expand or refine knowledge in the discipline. The ability to identify essential elements of complex problems and to formulate and support a critical generalization or hypothesis. Breadth and depth of knowledge of methods and procedures appropriate to research coupled with the ability to use the tools most relevant to the investigation to be undertaken. The ability to plan research adequately and to carry it to completion with care and precision, yet remaining flexible and sensitive to the possibility that data may indicate a need to modify preconceived hypotheses.

I WOULD EXPECT THIS STUDENT TO:

5

---- develop an original way of handling a research problem.

---- be familiar with the latest developments in his field.

4

---- be systematic in his gathering and ordering of data.

---- replicate previous research done by others with a carefully-conducted and well-reported experiment.

3

---- conduct a fairly routine and unexciting research project.

---- confine his attention to research matters of minor importance.

2

---- be unable to effectively apply a particular research technique.

---- attempt to carry out poorly planned research.

1

Degree of Confidence

☐ Very Confident ☐ Fairly Confident ☐ Not Very Confident

Teaching

General Definition

The ability to communicate concepts of students, ability to interest students in subject matter. Includes enthusiasm for teaching, willingness to spend time both in preparation for classes and out of class time with students.

I WOULD EXPECT THIS STUDENT TO:

5 ---- show imagination and originality in teaching a traditionally dull topic to an undergraduate class.

---- help slower students voluntarily.

4

take considerable pains to help undergraduates with their work even
3 ---- though his presentation of material in class is frequently poor.

---- spend class time doing routine exercises.

2 ---- complain about having to teach the introductory course.

---- teach a class in which his students seek help from other instructors.

badger students and be generally unsympathetic to legitimate requests
---- for time extensions or specific help.

1

Degree of Confidence

☐ Very Confident ☐ Fairly Confident ☐ Not Very Confident